Effect of nonresponse in Lithuanian travellers' survey

Marijus RADAVIČIUS (MII, VU)
e-mail: mrad@ktl.mii.lt

1. Introduction

This applied study is based on Lithuanian tourism survey data collected by the Statistics Lithuania. The survey has rather high nonresponse rate, about 30%, and estimators which do not take into account the nonrespose may yield considerable bias. The goal of the study is to propose an appropriate method for the nonresponse adjustment of the estimators and to evaluate the effect of the nonresponse.

2. The Lithuanian tourism survey data

The fieldwork of the survey is carried out at a specific period of year (one week of a quarter) in border crossing checkpoints where permanent residents of Lithuania returning from abroad are interviewed. The specific features of the survey are the following.

(a) The population surveyed consits of all Lithuanian travellers returning from abroad during the period of interest (quarter, year). However, only adult travellers are interviewed in the survey.

(b) The checkpoints can be classified into several groups according to a kind of transport they attend (road, train, plane, and boat) and, for the first kind of transport, additionally according to a neighbor country. The fieldwork is organized in 16 main checkpoints and it is assumed that the remaining checkpoints where the fieldwork of the survey is not organized are the typical representatives of their groups.

In addition to the survey data, the State Border Guard Service presents the information about the total number of the Lithuanian citizens arriving from abroad through each of the checkpoint during the corresponding quarter. According to this information the sampling rate is about 0.6%. Hence we have natural stratification of the population by the checkpoint crossed.

In each stratum, i.e., the checkpoint chosen for the survey, a cluster sample is withdrawn. A group of travellers travelling together (family) forms a cluster and only one adult member of the group/family is interviewed.

Thus, we can suppose that the sampling design of the survey is stratified (cluster) sample. In addition, since the sampling rate in each stratum is low (less than 1%), we can also assume without significant loss of accuracy that the samples in strata are drawn with replacements.

(c) The survey is non-homogeneous in two aspects. First, all travellers are divided into two categories: the travellers that have spent at least one night abroad (overnight travellers) and those who have not spent (one-day travellers). The first category of travellers is surveyed using the long version of the questionnaire while the short questionnaire is used for the second category.

Second, some questions in the questionnaire are individual, the others are concerned with a whole cluster (group of travellers). For instance, age, sex, overall appraisal of the travel, etc., are the individual features, whereas aims, visited countries, duration, expenditures are common for all members of a cluster (cluster features). In the later case one may assume that all members of a cluster are interviewed. This leads to proportional to size sampling scheme of clusters in the strata. As for individual questions, we have simple samples in the strata.

3. Response model

Let S and R denote the sample and the response set respectively. Set

$$p_s = P\{s \in R|S\}, \quad s \in S.$$

Usually logistic regression model (see, e.g., [5], classical references on logistic regression are [1, 2, 4]) with some auxiliary variable (vector) $z = z(s)$ as a predictor is used to describe the dependence of the response probability p_s on $s \in S$:

$$p_s = p(z(s)) = \frac{\exp\{\beta^T z(s)\}}{1 + \exp\{\beta^T z(s)\}},$$

where β is an unknown parameter to be estimated from the data.

The main assumption is that, for a given sample S and any pair $(s, r), s \neq r$, of respondents in the sample S, they both respond or nonrespond independently from each other [5]:

$$P\{s, r \in R|S\} = p_s p_r. \tag{1}$$

This assumption enables one to apply the weighted least squares method for the estimation of β. For the maximum likelihood method and testing of hypotheses, condition of the pairwise independence (1) should be replaced by the condition of the mutual independence.

An important for practice special case of this model is so-called RHG (*Response Homogeneity Group*) model [5]. As the name suggests, in this model the subsamples S_h,

$h = 1, \ldots, H$, of respondents with the same response probability form the whole sample S, i.e.,

$$p_s \equiv p(h), \quad s \in S_h, \ h = 1, \ldots, H.$$

Given response probabilities p_s, the adjusted for nonresponse estimators are obtained by replacing sampling design weights w_s with the weights $W_s = w_s/p_s$, $s \in S$.

3.1. *Unit nonresponse*

There are three variables available at the pre-sampling level, which can be applied to form subsamples S_h in the RHG model for the unit nonresponse: interviewer's identifier (I), checkpoint (C), and date (D). Let $U = U_s$ be an indicator of the unit respose, $U_s = 1$, if $s \in R$, and $U_s = 0$ otherwise.

The log-linear model fitted to the survey data is [*U I*] [*D C*] in usual log-linear modeling notation (see [1, 2, 4]). This means that pairs of variables (U, I) and (D, C) are independent. Hence, the unit nonresponse rate depends only on an interviewer and we obtain the RHG model with h being the interviewer's number.

3.2. *Item noresponse for expenditures*

Questions concerning expenditures are main source of the item noresponse. Logistic regression models are fitted separately for each quarter and each type of travellers, one-day and overnight. The fitted models include as predictors (explanation variables) interviewer's identifier, the purpose of the travel, the cluster size and other. However, only the first one, interviewer's identifier, is included in all the fitted logistic regression models and a simple model with this unique predictor has almost the same forecasting and goodness-of-fit characteristics as the former.

4. Estimation procedures

Let x and y denote variables of interest and let w_s denote the weight coefficient of a sample element $s \in S$. The total and the mean of x in a domain $D \subseteq U$ are estimated by the weighted sum $\widehat{X}_D$,

$$\widehat{X}_D = \sum_{s \in S \cap D} x_s w_s, \tag{2}$$

and the weighted mean $\hat{x}_D$,

$$\hat{x}_D = \frac{\sum_{s \in S \cap D} x_s w_s}{\sum_{s \in S \cap D} w_s}, \tag{3}$$

respectively. The estimator of the ratio of the totals of y and x, $\hat{\rho}_D$, is given by

$$\hat{\rho}_D = \hat{\rho}_D^{(x/y)} = \frac{\widehat{X}_D}{\widehat{Y}_D}. \tag{4}$$

The weight coefficients w_s are determined by sampling design and may additionally depend on the response probabilities and some auxiliary variables [3, 4]. Below we present a description of a procedure of calculating weights for the tourism survey.

Let g be a checkpoints' group index, $g = 1, \ldots, G, G$ being the total number of the groups (in fact $G = 7$), and let J_g be a set of identifiers of the all checkpoints belonging to the gth group and $J_{(g)}$ be the subset of J_g of those checkpoints wherein the fieldwork of the survey is organized. Further, let

$$N_g = \sum_{j \in J_g} N_j, \quad M_g = \sum_{j \in J_{(g)}} N_j, \quad n_g = \sum_{j \in J_{(g)}} n_j, \quad g = 1, \ldots, G,$$

where N_j is the total number of the travellers through the jth checkpoint during the period under consideration and n_j is the corresponding number of the interviewed travellers. Let $S_{(j)}$ denote the jth stratum of the sample, i.e., the set of all interviewed travellers crossing the jth checkpoint. Define

$$w_s^{(S)} = \frac{N_j N_g}{n_j M_g}, \text{ for } s \in S_{(j)}, \quad j \in J_{(g)},\ g = 1, \ldots, G. \tag{5}$$

Formula (5) gives the weights of sample elements in case of simple stratified sample.

REMARK 1. In case of (proportional to size) cluster sample in the strata the same weights are used, however mean value of the variable x over the cluster is substituted for x_s.

In view of the high rate of the nonresponse and other specific features of the survey discussed in (a)–(c), formula (5) should be modified. Since only the adult travellers are interviewed, the number N_j in formula (5) should be replaced by the total number $N_j^{(A)}$ of the *adult* travellers through the jth checkpoint. Hence an estimator of the proportions q_j of the adult travellers is needed. It is convenient to assume for a moment that only the adult travellers constitute the population of interest. Then the size k_s of the sth cluster is equal to the number a_s of the adults in the sth group of travellers. In view of Remark 1, the formulas (2) and (4) with $x_s \equiv 1$ and $y_s = m_s/a_s$, where m_s is the size of the sth group of travellers, yield

$$\hat{q}_j = \frac{n_j}{\sum_{s \in S_{(j)}} m_s/a_s}. \tag{6}$$

Thus, the new weights are

$$w_s^{(A)} = \frac{\hat{q}_j N_j N_g}{n_j M_g}, \quad s \in S_{(j)},\ j \in J_{(g)},\ g = 1, \ldots, G. \tag{7}$$

Below we present an adjusted for the nonresponse estimate of q_j (formula (10)).

As noticed in the previous section, one can assume that the unit nonresponse rate depends only on an interviewer's identifier h $(h = 1, \ldots, H)$ and the rate of the item nonresponse about the expenditures additionally depends on the traveller's type (one-day/overnight). Let $p_s^{(U)}$ and $p_s^{(i)}$ denote the estimated probabilities for the unit and the item response, respectively, of the sth traveller, $s \in R$. Set

$$p_s^{(E)} = p_s^{(U)} p_s^{(E)}, \quad s \in R. \tag{8}$$

Then the adjusted for the nonresponse weights

$$w_s^{(AR)} = \frac{\tilde{w}_s^{(A)}}{p_s} = \frac{\hat{q}_j^{(R)} N_j N_g}{n_j M_g p_s} \quad \text{for } s \in S_{(j)},\ j \in J_{(g)},\ g = 1, \ldots, G, \tag{9}$$

where $p_s = p_s^{(E)}$, if the expenditures are the variables of interest, and $p_s = p_s^{(U)}$ otherwise, $\tilde{w}_s^{(A)}$ is obtained from $w_s^{(A)}$(7) by substituting $\hat{q}_j^{(R)}$ for $\hat{q}_j$,

$$\hat{q}_j^{(R)} = \frac{\sum_{s \in S_{(j)} \cap R} 1/p_s^{(U)}}{\sum_{s \in S_{(j)} \cap R} m_s/(a_s p_s^{(U)})} \tag{10}$$

is the adjusted for the nonresponse estimate of q_j, the proportion of the adults in the jth stratum.

5. Results

The estimates for variables of interest are calculated using adjusted for the nonresponse weights $w_s^{(AR)}$ (9), simple (standard) weights $w_s^{(S)}$ (5) with the response set substituted for the sample, and certain intermediate weights. The first estimate is considered to be the best. For cluster (group) indices, common to whole group of travellers, the respective modifications outlined in Remark 1 are applied to the estimating procedures.

The effect of the nonrespose and other improvements of the standard estimating scheme is evaluated by comparison of the *relative corrected bias* with the *coefficient of variation* of the best estimate. The relative corrected bias is calculated as the ratio with the numerator equal to the bias of the estimate under consideration with respect to the best one and the denominator equals to the best estimate. As expected, the main improvement is achieved for expenditures. For the other indices, both individual and cluster, the relative corrected bias yields less than 10% of the coefficient of variation.

In the Table 1, figures obtained for the expenditures are presented. Additionally, an intermediate estimate, called "natural", is included. This estimate is based on the assumption that homogeneity groups for the unit response in RHG model coincide with the strata. This is the only difference of the "natural" estimate from the best one.

Table 1

Comparision of estimates for the expenditures

Quarter	Type	Coeff of Variation for Mean	Coeff of Variation for Total	Bias Simple Mean	Bias Simple Total	Bias Natural Mean	Bias Natural Total
2002 2	One-day	0.062	0.064	−0.006	0.069	−0.028	0.151
	Overnight	0.124	0.121	0.310	0.014	0.008	0.363
2002 3	One-day	0.218	0.219	−0.002	0.056	−0.034	0.053
	Overnight	0.076	0.077	0.123	−0.044	−0.038	0.365
2002 4	One-day	0.052	0.055	0.034	0.076	0.016	0.187
	Overnight	0.108	0.108	0.100	−0.002	−0.048	0.299
2003 1	One-day	0.069	0.072	0.029	0.061	0.018	0.149
	Overnight	0.160	0.159	0.153	0.011	−0.147	0.203

The first two columns of figures contain the coefficients of variation of the best estimate, the remaining four columns contain relative corrected bias of the estimates under the comparison. Note that usually the variation of the best estimates is much greater for overnight travellers with a striking exception in the third quarter of 2002. The "natural" estimates are rather poor for the total but not so bad for the mean. In contrary, the simple estimates behave better for the total. For the mean, their relative bias is close to or greater (up to 2.5 times) than the coefficient of variation of the best estimate, again with the exception in the third quarter of 2002. This suggests that the respondents of the different interviewers differ from each other not only in the unit nonresponse probabilities but in the expenditures of the travel as well.

References

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York (1990).

[2] R. Christensen, *Log-Linear Models*, Springer–Verlag, New York, Berlin (1990).

[3] S.L. Lohr, *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove (1999).

[4] Th.J. Santer, D.E. Duffy, *The Statistical Analysis of Discrete Data*, Springer–Verlag, New York, Berlin (1989).

[5] C.-E. Särndal, B. Swensson, J.Wretman, *Model Assisted Survey Sampling*, Springer–Verlag, New York (1992).

Neatsakymų įtaka Lietuvos keliautojų apklausoje

M. Radavičius

Šis taikomasis darbas remiasi Lietuvos turizmo apklausos duomenimis, surinktais Lietuvos Statistikos Departamente. Kadangi apklausoje buvo gana aukštas neatsakymų lygis, tai įvertiniai, kurie į tai neatsižvelgia, gali turėti didelį poslinkį. Darbo tikslas yra pasiūlyti tinkamą metodą neatsakymų sąlygotam įvertinių poslinkiui sumažinti bei įvertinti neatsakymų įtakos laipsnį ir pobūdį.